# Development Of An Advanced Image Classification Algorithm With Integrated Captioning

## SHORT SYNOPSIS

*Submitted in partial fulfillment of the requirement of the degree of*

## DOCTOR OF PHILOSOPHY

*to*

*Manav Rachna International Institute of Research and Studies*

*(Deemed to be University)*

*by*

**MONIKA CHAWLA**

**MRIIRSPHD 22/Ph.D/S/083**

Under the supervision of

**Dr. Rashmi Agrawal**

**Professor, School of Computer Applications, MRIIRS**

**School of Computer Application**

**Manav Rachna International Institute of Research & Studies**

**Sector-43, Surajkund–Delhi Road, Faridabad– 121004**

MANAV RACHNA
vidyanatariksha

## SCHOOL OF COMPUTER APPLICATIONS

### Certificate

I hereby declare that the Short Synopsis titled **"Development of an Advanced Image Classification Algorithm with Integrated Captioning"** being submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in the Department of Computer Application under School of Manav Rachna International Institute of Research & Studies, Faridabad during the academic year 2024. It is a bona fide record of original work carried out under the supervision of Dr Rashmi Agrawal and has not been submitted in full or part to any other university or institute for a degree or diploma. Further, it is certified that Short Synopsis is plagiarism free, and proof has been submitted.

Monika Chawla
MRIIRSPHD 22/Ph.D/S/083
monika.chawla12@gmail.com

Signature of the Supervisor

**Dr. Rashmi Agrawal**

**Professor, School of Computer Applications**

**(Chairperson, DRC)**

**(Dean, School of Computer Application)**

## Abstract

In computer vision, the most important task is image classification. Images are automatically classified into predefined classes based on the visual features extracted from them. The objective of image classification is to categorize images into one or more classes and examine patterns within the data. Techniques for image classification vary greatly from traditional approaches, relying on handcrafted features, such as those implemented through color histograms, edge detection, and texture analysis. However, with the development of classification algorithms utilizing deep learning techniques, current methods are primarily based on CNNs, which automatically learn hierarchical feature representations from large datasets. CNNs are particularly efficient for these tasks because they can capture spatial hierarchies and image patterns, such as edges, shapes, and textures, by applying convolutional, pooling, and fully connected layers. Other advanced techniques that have been intensively applied include transfer learning and ensemble methods, which involve using pre-trained networks or even combining the results of several models to further improve classification accuracy. These recent techniques have led to tremendous improvements in various applications of image classification tasks in terms of both performance and accuracy robustness. Building upon the image classification platform, this study develops an advanced algorithm that further classifies stronger methodologies to enhance performance.

Captioning images is a task that connects computer vision with natural language processing to generate descriptive textual captions for images. The goal is to produce sentences that coherently describe the visual content of an image, including the objects and their relationships within the scene. Traditional image captioning techniques rely on a combination of CNNs for extracting image features and RNNs, particularly LSTM

networks, for the generation of captions. Recently, transformer-based models have gained importance due to their enhanced capabilities in acquiring both visual and textual data at various levels of dependency across large distances. These techniques are crucial for enriching the semantic understanding of images. The implications of this make image captioning an integral component of several applications, such as automatic image description, accessibility tools, and multimedia retrieval systems.

In this study, the developed image captioning techniques are integrated with the proposed image classification algorithm to enhance contextual understanding and semantic representation of images. The primary objective of this research is to improve image classification capabilities. To begin, the existing techniques currently used in image classification and captioning will be critically analyzed for their weaknesses and challenges. Then, a new algorithm for image classification will be proposed and developed to enhance classification accuracy. Thus, integrating this captioning technique with the proposed classification algorithm can enhance contextual understanding and semantic representation of images. Finally, the proposed integrated model will be validated and tested using benchmark datasets to ensure its effectiveness and applicability in diverse scenarios.

*Keywords: image classification, image captioning, machine learning, deep learning, CNN, RNN and LSTM*

# Table of Contents


| | | |
|---|---|---|
| | Certificate | I |
| | Abstract | ii-iii |
| | Table of Contents | iv |

| Synod . | Contents | Page No. |
|---|---|---|
| 1 | Introduction<br>1.1 Image Classification<br>1.2 Image Captioning | 1- 3 |
| 2 | Literature Review | 4-12 |
| 3 | Research Gaps | 13 |
| 4 | Problem Statement | 13 |
| 5 | Research Objectives | 13 |
| 6 | Research Methodology | 14-16 |
| 7 | Expected Outcome of the Research | 16 |
| | References | 17-20 |
| | Similarity Verification | 21 |
| | Report Grammarly Certificate | |

# 1. Introduction

## 1.1 Image classification

It is the science of categorizing and labeling images into classes predefined by visual features. The feature extraction algorithms are based on a variety of machine learning or deep learning techniques that determine the features of an image and classify it into an assigned class. For instance, in an image classification task, one can use an algorithm that categorizes whether a cat, dog, or car is present in each image.

In computer vision, image classification is a crucial component because it enables machines to see and analyze visual information. This procedure requires extensive datasets composed of labeled images to train the models, allowing the algorithm to identify patterns, edges, textures, and several other significant characteristics. Widely used frameworks for this task include deep learning models, which are particularly proficient in identifying these patterns through multiple layers of filters and transformations. This capability facilitates precise classification across diverse categories, regardless of alterations in image scale, rotation, or translation.

As images are being generated in vast numbers by diverse sectors like healthcare, surveillance, autonomous driving, and social media, accurate and efficient image classification systems have become essential. Yet, the significant breakthroughs made through deep learning models—CNNs, ResNets (Residual Networks), DenseNet (Densely Connected Convolutional Networks), VGG (Visual Geometry Group Network), and Inception Network (Google Net)—remain open to further improvement. This highlights the semantic gap that exists between the visual content and its meaningful interpretation, as well as the complications encountered in the identification of images within complex contexts.

To overcome these challenges, this study aims to develop a model that simplifies the task of classification while integrating image captioning. Basic image captioning is generally regarded as the process of creating a sentence that describes what is in an image. This approach extends the context beyond what is achievable by traditional classifier-based methods that only classify based on visual perception. By combining image classification with natural language generation, the proposed method seeks to make image classification more accurate and understandable.

### 1.1.1 Image classification techniques

CNNs and advanced variants form the primary backbone of deep learning models for image classification, as shown in Figure 1. AlexNet pioneered deep learning with the introduction of ELU activation and dropout; subsequently, Genet developed deep networks using very small convolutional filters. GoogleNet, or Inception, created the Inception Module to facilitate efficient processing at different scales, while ResNet employs skip connections to allow for deeper networks. DenseNet optimizes connections by linking each layer with one another, increasing feature reuse. Later, lightweight models such as MobileNet and EfficientNet were developed, focusing deeply on efficiency and specific applications like mobile use cases. In a more recent study, ViT (Vision Transformers) applied transformer architectures from NLP to images. NASNet introduces an automated network architecture design that balances precision and efficiency.

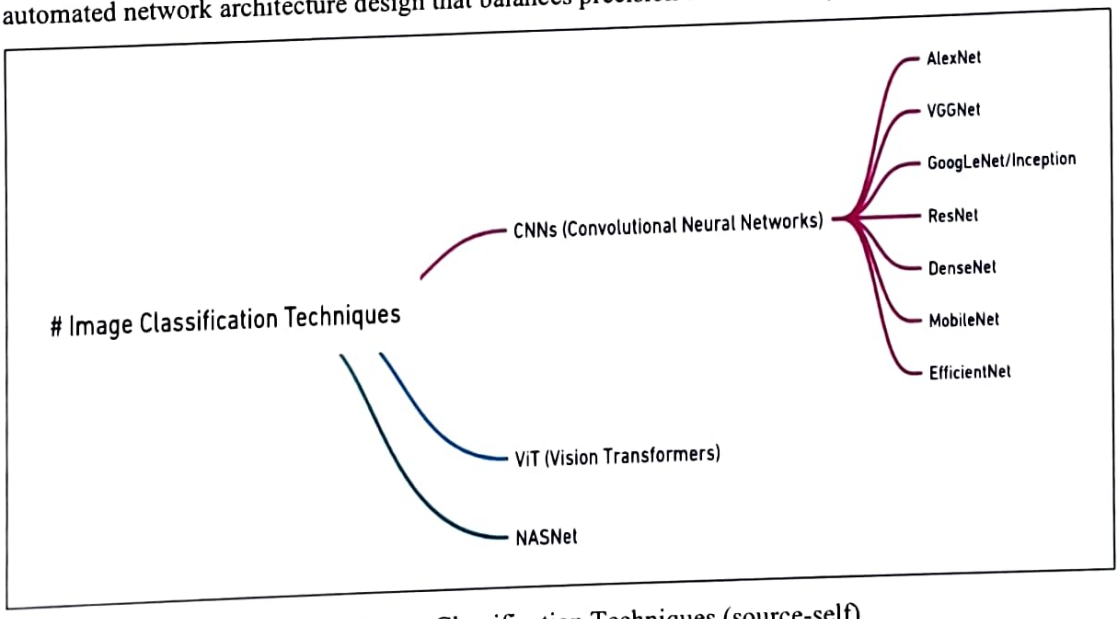


Figure 1- Image Classification Techniques (source-self)

## 1.2 Image caption

Its generation employs advanced deep learning techniques to analyze visual content and create original captions. This approach emphasizes generating unique, semantically accurate descriptions, providing more specificity than retrieval-based methods. Deep learning-based captioning techniques encompass a variety of learning approaches illustrated in Figure 2, including supervised, unsupervised, and

reinforcement learning. Several methodologies utilize an encoder-decoder structure, while others adopt a more complex compositional method that integrates attention mechanisms, semantic concepts, and various artistic elements. Although most deep learning models for generating image captions rely on long short-term memory (LSTM) networks, some utilize CNNs or RNNs. These deep learning techniques facilitate the generation of captions for entire scenes or specific elements within images, offering versatility and a diverse range of descriptive possibilities.

### 1.2.1 Image captioning techniques

The most prevalent techniques shown in Figure 2 involve combining CNNs with RNNs, especially LSTMs, to perform feature extraction on images and generate captions. Common architectures include the Encoder-Decoder framework, where attention is often added to allow the model to selectively focus on certain parts of the image while generating captions. Attention has further improved in the forms of Top-Down and Bottom-Up Attention. More recently, transformers have replaced RNNs in captioning. Multimodal models combine vision and text data, while GANs and reinforcement learning methods like SCST improve the quality of the captions produced. Dense captioning provides more descriptive content by giving multiple regions captions for an image.

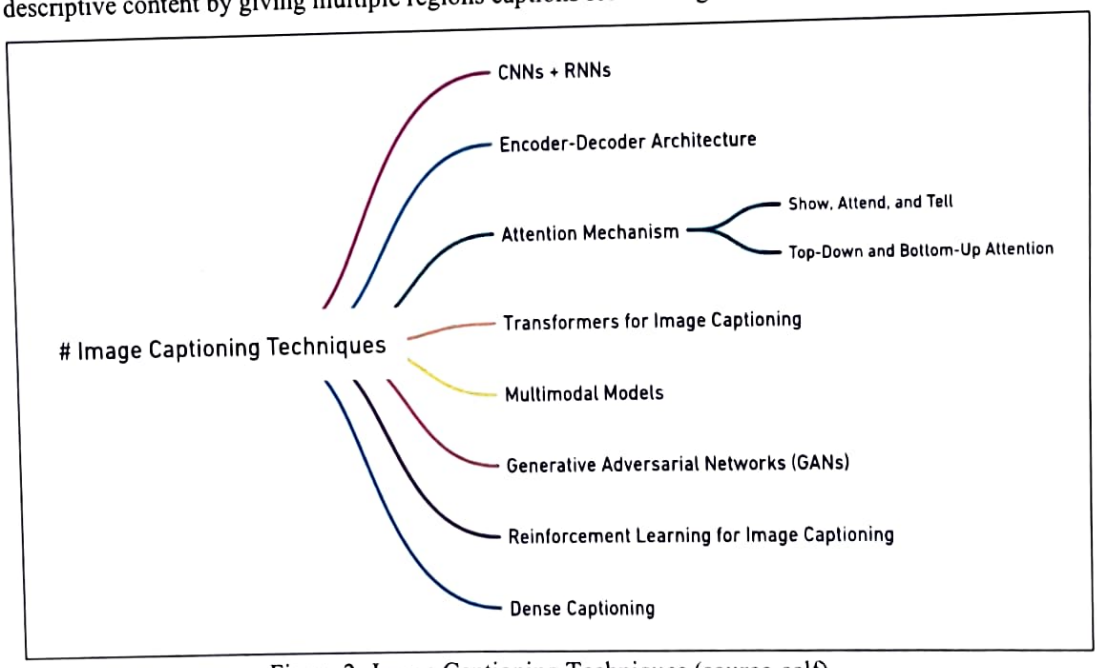


Figure 2- Image Captioning Techniques (source-self)

## 2. Literature Review

The literature covering image processing in the fields of image classification and image captioning covers so many decades that it explains how techniques have changed from conventional practices to more cutting-edge deep learning methods. This review is therefore a collation, in chronological order, of the most important works in this field and their implications for progress in Table 1.

Table 1- Image Processing Literature Review [Source : SELF]

| S. No. | Name of Author's | Method/ Technique | Result | Research Gap |
|---|---|---|---|---|
| 1 | K. Sanghvi, A. Aralkar, S. Sanghvi and I. Saha [4] | Convolutional Neural Network (CNN) for training a fauna image classifier, using tools like VGG16, TensorFlow, and Leaky ReLU | The result is that the proposed CNN-based method achieved 91.84% accuracy for fauna image classification. | A minimum of 1000 labeled training images and over 300 labeled validation images are required to train the model effectively for new animal classes. |
| 2 | A. Singh, P. Singh [1] | Convolutional Neural Networks (CNNs) for image classification, leveraging deep learning to automatically extract features and improve classification accuracy. | The model training process was optimized using early stopping to prevent overfitting, and the model's performance was tracked through validation accuracy. | Convolutional Neural Networks demand a large amount of labeled data and significant computational power, making the training process resource-intensive and challenging. |
| 3 | Chun-Fu (Richard) Chen, Quanfu Fan, Rameswar Panda [5] | The study proposes a dual-branch vision transformer model that processes image patches of different sizes separately to learn multi-scale feature representations for image classification, using a token fusion module based on cross-attention to reduce computational complexity. | The proposed approach outperformed recent models like DeiT by a significant margin of 2% on the ImageNet1K dataset, with only a moderate increase in FLOPs and model parameters, demonstrating better or comparable performance to other vision transformers and efficient CNN models. | The study does not fully explore the potential trade-offs of the increased model complexity and computational demands, particularly in resource-constrained environments, and lacks real-world testing beyond standard datasets like ImageNet1K. |
| 4 | H. Singh Gill, O. Ibrahim Khalaf, Y. Alotaibi, S. Alghamdi, and F. Alassery [6] | Multi-model deep learning approach that ensembles CNN, RNN, and LSTM algorithms for fruit recognition and classification. | A novel-fruit recognition classification strategy, namely the multi-model CNN-RNN-LSTM model. And represents the accuracy analysis of the multi-model classifier, RNN-96%, CNN-96.5%, and proposed- 97.7% | A single feature descriptor cannot adequately address complex classification problems, requiring more advanced and efficient feature fusion methods to manage the diverse characteristics of fruits effectively. |
| 5 | T. Sharma, R. Nair, and S. Gomathi [7] | CNN-based model for supervised classification, on the breast cancer picture dataset | The deep learning model using a Convolutional Neural Network with transfer learning achieved | Smaller batch sizes accelerate training and also give better early learning but do not guarantee |

| | | | | |
|---|---|---|---|---|
| | | | an accuracy of 98.4% in breast cancer analysis on medical images | convergence to the global optimum. |
| 6 | Ankit Yadav, Ibtesaam Rais, Manoj Kumar, Anurag Sharma and Abhay Kushwaha [8] | Deep Neural Network (DNN) implemented with TensorFlow for image classification, focusing on model training, image categorization, and managing accuracy and overfitting through epochs. | Deep Neural Networks (DNNs) have low accuracy for image tasks, Convolutional Neural Networks (CNNs) provide more optimized results with 90% | achieved up to 90% accuracy in classifying flower images using a DNN with TensorFlow, the results might still be constrained by the risk of overfitting and may require further tuning |
| 7 | Hee E. Kim, Alejandro Cosa- Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E. Maros and Thomas Ganslandt [9] | Achieving optimal performance may require careful incremental fine-tuning and learning rate adjustments, | Studies evaluated multiple models, with deep models like Inception being the most common, and in transfer learning, feature extraction and fine-tuning from scratch were the preferred approaches, while hybrid methods were less frequently used. | The limitation is that while transfer learning is effective with limited data, relying on deep models as feature extractors may still require careful tuning to maintain predictive power without sacrificing accuracy. |
| 8 | Gizem ¨Ozerol ¨Ozman a,, Semra Arslan Selçuk a, Abdussamet Arslan [10] | This study applied CNN-based damage assessment algorithms, including AlexNet, VGG19, and ResNet50, to classify building damage categories ("Inspected, Restricted Use, Unsafe") from exterior images in post-earthquake damage assessments. | The study found that VGG19 produced results validation accuracy of 79% in 10 hours, while AlexNet offered 55 % in 1.23 min, and ResNet50 with 77% in 5 minutes. | The study's limitations include longer training times for deeper networks like VGG19 and potential overfitting in AlexNet algorithms, which affects the accuracy and loss values. |
| 9 | R. Archana, P. S. Eliahim Jeevaraj [11] | Image denoising using Self2Self NN, DnCNNs, and DFT-Net; image enhancement techniques like Novel RetinexDIP, Unsharp Masking, and LE-net; segmentation using foundational and advanced models; feature extraction with CNNs and LSTM-augmented CNNs; and image classification using architectures like Residual Networks and CNN-LSTM. | Such a survey presents and compares various deep learning techniques for image denoising, enhancement, segmentation, feature extraction, and classification, showcasing their strengths, challenges, and applicability across various domains- medical, satellite, and botanical imagery. | Deep learning methods face challenges like detail loss, handling complex scenes, robustness, computational complexity, data dependency, and interpretability. |
| 10 | Jatinder Kumar, Surya Narayan Panda, and Devi Dayal [12] | The study explores the use of machine learning, specifically deep learning algorithms, for feature extraction and image segmentation in ultrasound imaging of thyroid disorders. | Deep learning algorithms successfully recognize and quantify patterns in clinical ultrasound images, improving the analysis and classification of thyroid-related illnesses. | The study highlights certain research challenges and the need for future enhancements, particularly in addressing algorithm optimization, generalization, and the need for diverse datasets. |
| 11 | Y. Gong, H. Zhang, R. Xu, Zhou Yu, and J. Zhang [13] | The study aims to develop an accurate colorectal polyp detection model by analyzing and leveraging contemporary | The proposed deep learning model significantly enhances the precision and efficiency of | The research suggests a need for further exploration of deep neural network technology to optimize |

| | | | | |
|---|---|---|---|---|
| | | deep neural network-based target detection algorithms to improve feature learning and representation. | colorectal polyp detection, aiding physicians in improving diagnostic accuracy and simplifying the diagnostic process. | accuracy and efficiency, but it doesn't address specific challenges such as model generalization, dataset diversity, or potential improvements in real-world clinical settings. |
| 12 | Yue, Y., & Li, Z. [14] | The study proposes MedMamba, a novel hybrid model for medical image classification that integrates convolutional layers with state space models (SSMs) to capture both local features and long-range dependencies. The model uses a new SS-Conv-SSM block, employing grouped convolution and channel-shuffle strategies to reduce computational burden and model parameters. | MedMamba demonstrated competitive performance across 16 datasets, including 10 different imaging modalities and 411,007 images, outperforming many state-of-the-art methods in medical image classification tasks. | The study does not thoroughly address the potential limitations in real-world deployment, particularly in terms of scalability across other complex medical imaging modalities, and further exploration is needed to optimize SSM-based AI algorithms for more practical applications in healthcare settings. Additionally, potential limitations in clinical adoption or resource-constrained environments are not deeply discussed. |
| 13 | J. Wang, X. Li, Y. Jin, Y. Zhong, K. Zhang, C. Zhou [15] | The study uses deep neural networks to identify human behaviors by integrating bone point data from Microsoft Kinect cameras with conventional images, evaluated on the MSR3D dataset. | The algorithm achieved consistently high accuracy in behavior recognition, demonstrating reliability across various scenarios. It also substantially improved the accuracy of pedestrian behaviour detection in video footage. | The study does not address potential challenges such as scalability to more complex datasets, generalization to diverse real-world environments, or the performance impact in situations with varying camera angles and lighting conditions. Further research is needed to test the algorithm's robustness in such varied scenarios. |
| 14 | R. Khanam, M. Hussain, R. Hill, and P. Allen. [16] | The study reviews the use of Convolutional Neural Networks (CNNs) in industrial applications for surface defect detection, focusing on object detection methodologies and suitable hardware platforms to deploy CNN-based architectures in various scenarios like pallet racks and display screens. | CNNs have shown high accuracy and real-time performance in automating defect detection across different surfaces, making significant improvements in quality inspection processes in industrial applications. | The study identifies challenges such as limited datasets, high computational complexity, and domain-specific nuances that need further research for broader and more efficient application of CNNs in industrial defect detection. |
| 15 | S. Li, P. Kou, M. Ma, H. Yang, S. Huang, And Z. Yang [17] | The study proposes a new semi-supervised deep learning model that replaces traditional neurons with quadratic neurons and uses quadratic convolution layers to improve feature extraction. | The proposed method improves the generalization performance and image classification accuracy of the model, as demonstrated by experiments on two large medical datasets. | The study highlights challenges such as the scarcity of labelled medical image data, large image sizes, and image quality instability, which affect the model's robustness and generalization, indicating |

| | | | | |
|---|---|---|---|---|
| | | | | the need for further research to address these issues. |
| 16 | Z. He, C. Su and Y. Deng [18] | The study develops Deformable Oriented YOLOv4 (DO-YOLOv4) for crack detection in concrete structures, enhancing feature extraction with deformable convolutional layers and enclosing crack regions with oriented bounding boxes. The method is then combined with image processing techniques. | The results show that the DO-YOLOv4-IPTs achieves superior segmentation accuracy with less labelling effort, outperforming traditional CNN-based crack segmentation methods, especially in the presence of background noise. | The study does not address potential challenges in real-world deployment, such as scalability to different structural materials or conditions, and further research is needed to explore how the model performs across diverse environments and crack types. Additionally, improving the labelling process further could enhance the practicality of the method. |
| 17 | Y. Jacob Sandvik, C. M. Futsæther, K. Hovde Liland and O. Tomic [19] | The literature review analysed 33 studies from four scientific databases to assess The effectiveness of image-processing methods and machine-learning algorithms in computer vision for grading and scaling wood logs, categorizing studies by goals (e.g., log end grading, log side grading, log segmentation) and comparing inputs, models, performance, and autonomy levels. | The review found that deep learning models performed better in log pile scaling and log segmentation, with an increased use of cameras over laser scanners. | The review highlights the challenge of comparing model performance due to varying goals and metrics across studies, and further research is needed to standardize evaluation criteria to facilitate more effective comparison and advancement in wood log grading and scaling technologies. |
| 18 | R. Wang, Y. Nie, and J. Geng [20] | The study proposes a multiscale superpixel-guided weighted graph convolutional network (MSGWGCN) for classifying Polarimetric Synthetic Aperture Radar (PolSAR) images. The model integrates multiscale superpixel features into a weighted graph convolutional network to better utilize land cover information and fuses pixel-level features using a multiscale feature cascade fusion module to preserve classification details. | Experiments conducted on three PolSAR datasets demonstrate that the proposed MSGWGCN outperforms other advanced methods in PolSAR image classification, particularly in addressing misclassification at the boundaries of different land cover objects. | The research does not thoroughly address the computational complexity or scalability of the model when applied to larger or more diverse datasets, indicating a need for further research into optimizing the approach for broader real-world applications. |
| 19 | Benazir Meerasha, Martin Sagayam [21] | The study develops and tests an object-based (OB) Land Use–Land Cover (LULC) classification approach in Google Earth Engine (GEE) by combining the Random Forest (RF) machine learning algorithm for classification, | This method was successfully tested on a 304 km² study area in Telangana, India, using Sentinel 2 (S2) data, demonstrating the effectiveness of combining OB methods with ML | The research highlights the challenges of object-based classification in GEE, particularly in adjusting parameters and overcoming computational limitations. Further research is needed to refine OB methods within |

| | | | | |
|---|---|---|---|---|
| | | the Gray-Level Co-occurrence Matrix (GLCM) for calculating texture indices, and the Simple Non-Iterative Clustering (SNIC) algorithm to identify spatial clusters. Principal Component Analysis (PCA) is applied to combine the GLCM texture indices into a single band for OB classification. | algorithms and texture analysis for LULC classification. | GEE for more efficient processing and broader application. |
| 20 | R. Archana, P. S. Eliahim Jeevaraj [22] | The study reviews a wide array of image processing techniques, including traditional methods and deep learning (DL) models, covering tasks such as denoising, enhancement, segmentation, feature extraction, and classification. It highlights specific methods like Denoising CNNs, Mask-RCNN, CNN, and Residual Networks, comparing their approaches to image analysis and manipulation. | The review highlights the strengths of DL models in learning feature representations directly from data, providing improved performance in tasks like denoising, segmentation, and classification. Traditional methods rely on predefined algorithms, while DL models excel at automating feature extraction and achieving higher precision in complex tasks. | Challenges identified include the computational demands and complexity of DL models, difficulties with robustness in real-world scenarios, data augmentation, parameter tuning, and the trade-offs in interpretability and performance across various image processing tasks. Further research is needed to address these issues, particularly in optimizing computational resources and enhancing model robustness. |
| 21 | Xiongquan Li, XukangWang, Xuhesheng Chen, Yao Lu, Hongpeng Fu & Ying Cheng Wu [23] | This paper introduces and evaluates new selection strategies in the Active Learning framework: Similarity-Based Selection, Prediction Probability-Based Selection, and Competence-Based Active Learning. The goal is to find the most informative images from unlabelled datasets for image classification. The methods are tested on popular datasets like Cifar10 and Cifar100, with the objective of minimizing the required training data while maximizing model performance. | The proposed selection methods outperformed random selection and conventional techniques, demonstrating superior performance in identifying informative data samples and enhancing the Active Learning process for image classification tasks. | The study does not address challenges related to scalability across more complex datasets or real-world applications, and further research is needed to optimize the Active Learning process in terms of computational efficiency and model generalization in diverse data environments. |
| 22 | A. Senokosov, A. Sedykh, A. Sagingalieva, B. Kyriacou, and A. Melnikov [24] | The study introduces two quantum machine learning models for image classification. The first model is a hybrid quantum neural network with parallel quantum circuits, designed to function effectively in the noisy intermediate-scale | The first model achieved 99.21% accuracy on the MNIST dataset, outperforming existing quantum-classical models with significantly fewer parameters. It also demonstrated over 99% accuracy on the Medical | While the models show promising results, the study does not thoroughly address challenges in scaling these quantum models for larger and more complex datasets, or in deploying quantum machine learning systems in practical, real-world |

| | | | | |
|---|---|---|---|---|
| | | quantum (NISQ) era. The second model includes a hybrid quantum neural network with a Quanvolutional layer to reduce image resolution via a convolution process. Both models aim to enhance computational efficiency and accuracy in image classification. | MNIST dataset and 82% accuracy on CIFAR-10, showing strong generalizability. The second model matched the performance of its classical counterpart while using four times fewer trainable parameters and surpassed classical models with equal weight parameters. | environments where quantum hardware limitations may pose significant barriers. Further research is needed to explore these aspects. |

## 3. Research Gaps:

i. Currently, there are no integrated models that efficiently execute image classification with captioning within one framework.

ii. Complex scenes in context are not captioned by the current models with any accuracy or meaning.

iii. The current algorithms for image classification and captioning do not meet transform-invariance, such as scaling, rotation, and occlusion.

iv. Existing algorithms demonstrate restricted generalization capabilities across a variety of domains and intricate image scenes.

v. Many models overfit badly especially when used on small datasets or special ones.

vi. Generated captions still lack interpretability and hence indirectly impact their relevance and contextual accuracy.

vii. Handling Multi-object or ambiguous scenes while keeping the classification and captioning contexts together remains a challenge.

## 4. Problem Statement:

The current separation of image classification and captioning tasks results in inefficiencies, particularly in real-time applications and resource-constrained environments. Existing models struggle to generate accurate and informative captions for complex scenes involving transformations like scaling, rotation, or occlusion. This research aims to design an integrated algorithm that enhances classification accuracy while generating meaningful and contextual captions.

## 5. Research Objectives

i. To critically analyze existing image classification and image captioning techniques

ii. To design and develop a novel image classification algorithm to enhance the performance of image classification.

iii. To integrate image captioning techniques within the developed classification algorithm, improving the contextual understanding and semantic representation of images

iv. To validate and test the proposed integrated model on benchmark datasets.

## 6. Research Methodology

The research methodology is structured and multi-phased with each phase designed to work in a systematic way to evolve and test an advanced image classification algorithm integrated with a captioning system. The objectives of each phase are outlined as data collection and preprocessing, model training, optimization, testing, and refinement.

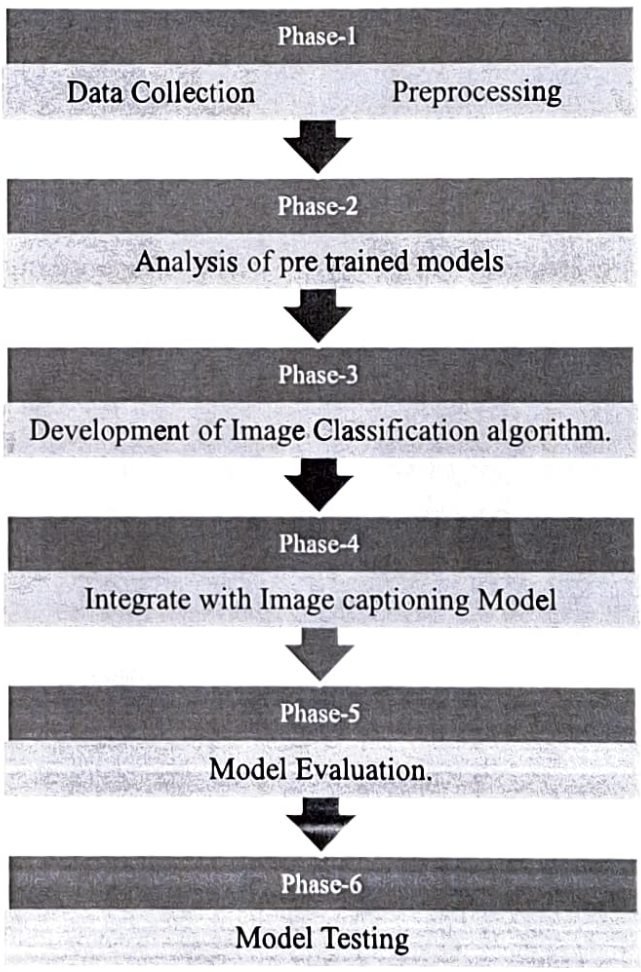


Figure 3- Phases of Proposed Research (source-self)

**Phase 1: Data collection and preprocessing:** would be initiated in the form of gathering the required datasets for image classification. Quite well-known datasets like ImageNet and CIFAR10 would be used for this purpose. Suitable data augmentation like scaling, rotation, and translation would be applied to enrich the dataset so that the model learns to be robust against the transformations of images.

**Phase 2: Analysis of already used pre-trained Model:** This demands an assessment of the pre-trained models that were already used to determine the best-performing algorithms that can be adopted in image classification tasks. Models like CNNs, ResNet (Residual Networks), DenseNet (Densely Connected Convolutional Networks), VGG (Visual Geometry Group Network), Inception Network (GoogleNet), etc, as well as Inception. Scaling, rotations, as well as any form of translations, will be adopted in images and assessed as to which one is the best-suited model for fine-tuning with the preprocessed datasets upon which a classification model can be based.

**Phase 3: Development of the Image Classification Algorithm:** The development of a strong classification model, using a deep learning model as the base, will be emphasized with optimization to serve for transformed images effective classification. Images will be preprocessed by resizing and normalizing; associated text labels will be given to the image.

**Phase 4: Integrate with Image Captioning Model:** this will consist of using a deep learning-based captioning algorithm to come up with meaningful descriptions of each classified image. It will be trained in pairs of images and texts, where the quality of the captions generated will be appraised by standard metrics.

**Phase 5: Model Evaluation and Testing:** The Hybrid model will be intensively tested on benchmark datasets. The standard performance metrics including accuracy, precision, recall, and the F1-score will be used to test the image classification model. The captioning module is evaluated using BLEU and ROUGE scores and objective human evaluation. Testing in this phase will also check how well the model tolerates variations in images such as scaling, rotation, and translation.

## 7. Expected Outcome of the Research

The expected outcome of the study is a very sophisticated image classification algorithm that provides integrated captioning of images but significantly boosts classification accuracy while generating textually relevant and meaningful descriptions. The algorithm will be designed to run efficiently on computational platforms and to be robust. In addition, the model is expected to be robustly generalizing across domains, incorporate improved feature fusion methodologies, and address issues of ethics, such as bias mitigation. Overall, this research is envisioned to bring about more scalable, interpretable, and ethically responsible solutions for the task of image classification and provide captioning also.

## References


[1] Ankita Singh, Pawan Singh, Image Classification: A Survey, Journal of Informatics Electrical and Electronics Engineering, 2020, Vol. 01, Iss. 02, S. No. 002, pp. 1-9 ISSN (Online): 2582-7006

[2] Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, *28*(5), 823-870.

[3] Olaode, A., Naghdy, G., & Todd, C. (2014). Unsupervised classification of images: A review. *International Journal of Image Processing*, *8*(5), 325-342.

[4] Sanghvi, Kavish and Aralkar, Adwait and Sanghvi, Saurabh and Saha, Ishani, A Survey on Image Classification Techniques (November 25, 2020). Available at SSRNhttp://dx.doi.org/10.2139/ssrn.3754116

[5] Chen, C., Fan, Q., & Panda, R. (2021). CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 347-356.

[6] Harmandeep Singh Gill, Osamah Ibrahim Khalaf, Youseef Alotaibi, Saleh Alghamdi and Fawaz Alassery, In Multi-Model CNN-RNN-LSTM Based Fruit Recognition and Classification, 2022, Tech Science Press, Intelligent Automation & Soft Computing

[7] Tripti Sharma, Rajit Nair, and S. Gomathi, Breast Cancer Image Classification using Transfer Learning and Convolutional Neural Network,In International Journal of Modern Research, Vol. 2, No. 1, 8-16, 2022.

[8] Ankit Yadav, Ibtesaam Rais, Manoj Kumar, Anurag Sharma and Abhay Kushwaha, Image Classification using Deep Learning and Tensorflow, In International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538, Volume 10 Issue V May 2022, Doi: https://doi.org/10.22214/ijraset.2022.43385

[9] Hee E. Kim , Alejandro Cosa-Linan, Nandhini Santhanam, Mahboubeh Jannesari, Mate E. Maros and Thomas Ganslandt, In Transfer learning for medical image classification: a literature review, Kim et al. BMC Medical Imaging (2022) 22:69 https://doi.org/10.1186/s12880-022-00793-7

[10] Gizem ¨Ozerol ¨Ozman a,, Semra Arslan Selçuk a, Abdussamet Arslan, Image classification on Post-Earthquake damage assessment:A case of the 2023 Kahramanmaras, earthquake, in Engineering Science and Technology, an International Journal,2023

[11] Homepage www.elsevier.com/locate/jestch

[12] R. Archana, P. S. Eliahim Jeevaraj, Deep learning models for digital image processing: a review, in Artificial Intelligence Review (2024) 57:11

[13] International Journal of Innovative Research in Computer Science and Technology (IJIRCST), ISSN(Online): 2347-5552, Volume-12, Issue-2, March 2024, https://doi.org/10.55524/ijircst.2024.12.2.14, Article ID IJIRD-1279, Pages 81-86.

[14] Yue, Y., & Li, Z. (2024). Medmamba: Vision mamba for medical image classification. arXiv preprint arXiv:2403.03849.

[15] Wang, J., Li, X., Jin, Y., Zhong, Y., Zhang, K., & Zhou, C. (2024). Research on image recognition technology based on multimodal deep learning. arXiv preprint arXiv:2405.03091.

[16] Rahima Khanam , Muhammad Hussain , Richard Hill , And Paul Allen., A Comprehensive Review of Convolutional Neural Networks for Defect Detection in Industrial Applications, IEEE Access,VOLUME 12, 2024,P-94250-94295; doi: 10.1109/ACCESS.2024.3425166.

[17] Sai Li, (Member, Ieee), Peng Kou, Miao Ma, Haoyu Yang, Shuo Huang, And Zhengyi Yang, Application of Semi-Supervised Learning in Image Classification: Research on Fusion of Labeled and Unlabeled Data, in IEEE Access,VOLUME 12, 2024,P-27331-27343; doi: 10.1109/ACCESS.2024.3367772.

[18] He, Z.; Su, C.; Deng, Y. A Novel Hybrid Approach for Concrete Crack Segmentation Based on Deformable Oriented-YOLOv4 and Image Processing Techniques. Appl. Sci. 2024, 14, 1892. https://doi.org/10.3390/app14051892

[19] Sandvik, Y.J.; Futsæther, C.M.; Liland, K.H.; Tomic, O. A Comparative Literature Review of Machine Learning and Image Processing Techniques Used for Scaling and Grading of Wood Logs. Forests 2024, 15, 1243. https://doi.org/10.3390/f15071243

[20] Ru Wang, Yinju Nie , and Jie Geng, Multiscale Superpixel-Guided Weighted Graph Convolutional Network for Polarimetric SAR Image Classification, IEEE Journal of Selected Topics In Applied Earth Observations And Remote Sensing, Vol. 17, 2024, Pg-3727-3741,

[21] Benazir Meerasha, Martin Sagayam, A Comparison of Pixel Based and Object Based Image Classification for Cropland Area Estimation, J. Electrical Systems 20-7s (2024): 2314-2322

[22] R. Archana, P. S. Eliahim Jeevaraj, Deep learning models for digital image processing: a review, Artifcial Intelligence Review (2024) 57:11 https://doi.org/10.1007/s10462-023-10631-z

[23] Xiongquan Li, XukangWang, Xuhesheng Chen, Yao Lu, Hongpeng Fu & Ying Cheng Wu, Unlabeled data selection for active learning in image classification, Scientifc Reports,(2024) 14:424 Doi: https://doi.org/10.1038/s41598-023-50598-z

[24] Senokosov, A., Sedykh, A., Sagingalieva, A., Kyriacou, B., & Melnikov, A. (2024). Quantum machine learning for image classification. Machine Learning: Science and Technology, 5(1), 015040.

[25] Jatinder Kumar, Surya Narayan Panda, and Devi Dayal, 36 Pediatric thyroid ultrasound image classifcation using deep learning: A review, Scientifc Reports,(2024) 14:424.

[26] Rathi, D., Jain, S., & Indu, S. (2017, December). Underwater fish species classification using convolutional neural network and deep learning. In *2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR)* (pp. 1-6). IEEE.

[27] Sanghvi, K., Aralkar, A., Sanghvi, S., & Saha, I. (2020). Fauna Image Classification using Convolutional Neural Network. *International Journal of Future Generation Communication and Networking*, *13*(1s), 08-16.

[28] Sun, X., Liu, L., Wang, H., Song, W., & Lu, J. (2015, December). Image classification via support vector machine. In *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)* (Vol. 1, pp. 485-489). IEEE.

[29] Diaz, R. A. N., Swandewi, N. N. T., & Novianti, K. D. P. (2019, August). Malignancy Determination Breast Cancer Based on Mammogram Image With K-Nearest Neighbor. In *2019 1st International Conference on Cybernetics and Intelligent System (ICORIS)* (Vol. 1, pp. 233-237). IEEE.

[30] Cai, Y. L., Ji, D., & Cai, D. (2010, June). A KNN Research Paper Classification Method Based on Shared Nearest Neighbor. In *NTCIR* (pp. 336-340).

[31] Zawbaa, H. M., Hazman, M., Abbass, M., & Hassanien, A. E. (2014, December). Automatic fruit classification using random forest algorithm. In *2014 14th International Conference on Hybrid Intelligent Systems* (pp. 164-168). IEEE.

[32] Shaha, M., & Pawar, M. (2018, March). Transfer learning for image classification. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 656-660). IEEE.

[33] Feng Chen, Xinyi Li, Jintao Tang, Shasha Li and Ting Wang (2021), A Survey on Recent Advances in Image Captioning. IOP Publishing IOP Publishing Journal of Physics: Conference Series (2021) 012053IOP Publishingdoi:10.1088/1742-6596/1914/1/0120532

[34] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D, Show and tell: A neural image caption generator, in *CVPR* 2015, pp. 3156-3164.

[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, *Computer Science*, pp. 2048–2057, 2015.

[36] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, Attention on attention for image captioning, in *ICCV* 2019, pp. 4634–4643.

[37] Aarthi R, Rishma G (2023) A Vision based approach to localize waste objects and geometric features exaction for robotic manipulation. Int Conf Mach Learn Data Eng Procedia Comput Sci 218:1342–1352. https:// doi. org/ 10.1016/j. procs. 2023. 01.113

[38] Abdar M, Samami M, Mahmoodabad SD, Doan T, Mazoure B, Hashemifesharaki R, Liu L, Khosravi A, Acharya UR, Makarenkov V, Nahavandi S (2021) Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. Comput Biol Med 135:104418. https:// doi. org/ 10. 1016/j. compb iomed. 2021.104418

[39] Aggarwal A, Kuma M (2020) Image surface texture analysis and classification using deep learning. Multimed Tools Appl 80(1):1289–1309. https:// doi. org/ 10. 1007/ s11042- 020- 09520-2

[40] Ahammad SH, Rajesh V, Rahman MZU, Lay-Ekuakille A (2020) A hybrid CNN-based segmentation and boosting classifier for real time sensor spinal cord injury data. IEEE Sens J 20(17):10092–10101. https:// doi. org/ 10. 1109/ jsen. 2020. 2992879

[41] Ahmad S, Ullah T, Ahmad I, Al-Sharabi A, Ullah K, Khan RA, Rasheed S, Ullah I, Uddin MN, Ali MS (2022) A novel hybrid deep learning model for metastatic cancer detection". Comput Intell Neurosci 2022:14. https:// doi. org/ 10.1155/ 2022/ 8141530

[21] Y. Pan, T. Yao, Y. Li, and T. Mei, X-linear attention networks for image captioning, in *CVPR* 2020, pp. 10 971–10 980.

[22] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, Meshed-memory transformer for image captioning, in *CVPR* 2020, pp. 10 578–10 587.

[23] J. Ji, Y. Luo, X. Sun, F. Chen, G. Luo, Y. Wu, Y. Gao, and R. Ji, improving image captioning by leveraging intra-and inter-layer global representation in transformer network, in *AAAI* 2021.

[24] R. Chatterjee, M. Weller, M. Negri, and M. Turchi, Exploring the planet of the apes: a comparative study of state-of-the-art methods for mt automatic post-editing, in *ACL* (Volume 2: Short Papers), 2015, pp. 156–161.

[25] L. Guo, J. Liu, S. Lu and H. Lu, Show, tell, and polish: Ruminant decoding for image captioning, in *TMM*, vol. 22, no. 8, pp. 2149–2162, 2019.

[42] Ahmed I, Ahmad M, Khan FA, Asif M (2020) Comparison of deep-learning-based segmentation models: using top view person images". IEEE Access 8:136361–136373. https:// doi. org/ 10. 1109/ access. 2020. 30114 06

[43] Aish MA, Abu-Naser SS, Abu-Jamie TN (2022) Classification of pepper using deep learning. Int J Acad Eng Res (IJAER) 6(1):24–31.

Ashraf H, Waris A, Ghafoor MF et al (2022) Melanoma segmentation using deep learning with testtime augmentations and conditional random fields. Sci Rep 12:3948. https:// doi. org/ 10. 1038/ s41598- 022- 07885-y

[44] Bouteldja N, Klinkhammer BM, Bülow RD et al (2020) Deep learning based segmentation and quantification in experimental kidney histopathology. J Am Soc Nephrol. https:// doi. org/ 10. 1681/ ASN. 20200 50597

[45] Cheng G, Xie X, Han J, Guo L, Xia G-S (2020) Remote sensing image scene classification meets deep learning: challenges, methods, benchmarks, and opportunities. IEEE J Select Topics Appl Earth Observ Remote Sens 13:3735–3756. https:// doi. org/ 10. 1109/ JSTARS. 2020. 30054 03

[46] Devulapalli S, Potti A, Rajakumar Krishnan M, Khan S (2021) Experimental evaluation of unsupervised image retrieval application using hybrid feature extraction by integrating deep learning and handcrafted techniques. Mater Today: Proceed 81:983–988. https:// doi. org/ 10. 1016/j. matpr. 2021. 04. 326

[47] Dey S, Bhattacharya R, Malakar S, Schwenker F, Sarkar R (2022) CovidConvLSTM: a fuzzy ensemble model for COVID-19 detection from chest X-rays. Exp Syst Appl 206:117812. https:// doi. org/ 10. 1016/j. eswa. 2022. 117812

[48] Gao C, Zhou J, Wong WK, Gao T (2019) Woven Fabric Defect Detection Based on Convolutional Neural Network for Binary Classification. In: Wong W (ed) Artificial Intelligence on Fashion and Textiles AITA 2018 Advances in Intelligent Systems and Computing. Springer, Cham. https:// doi. org/ 10. 1007/ 978-3- 319- 99695-0_ 37

[49] Gao X, Zhang M, Luo J (2022) Low-light image enhancement via retinex-style decomposition of denoised deep image prior. Sensors 22:5593. https:// doi. org/ 10. 3390/ s2215 5593

[50] Gill HS, Murugesan G, Mehbodniya A, Sajja GS, Gupta G, Bhatt A (2023) Fruit Type Classification using Deep Learning and Feature Fusion. Comput Electronic Agric 211:107990 https:// doi. org/ 10. 1016/j. compag. 2023. 107990

[51] Gite S, Mishra A, Kotecha K (2022) Enhanced lung image segmentation using deep learning. Neural Comput and Appl. https:// doi. org/ 10. 1007/ s00521- 021- 06719-8

[52] Hasti VR, Shin D (2022) Denoising and fuel spray droplet detection from light-scattered images using deep learning. Energy and AI 7:100130. https:// doi. org/ 10. 1016/j. egyai. 2021. 100130

[53] Hedayati R, Khedmati M, Taghipour-Gorjikolaie M (2021) Deep feature extraction method based on ensemble of convolutional auto encoders: Application to Alzheimer's disease diagnosis. Biomed Signal Process Control 66:102397. https:// doi. org/ 10. 1016/j. bspc. 2020. 102397

[54] Hussain E, Hasan M, Hassan SZ, Azmi TH, Rahman MA, Parvez MZ (2020) [IEEE 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)Kristiansand,15th IEEE Conference on Industrial Electronics and Applications (ICIEA) -Deep Learning Based Binary Classification for Alzheimera Disease Detection using Brain MRI Images. pp. 1115–1120. https:// doi. org/ 10. 1109/ iciea 48937. 2020. 92482 13.

in *ECCV*. Springer, 2016, pp. 382–398.

MANAV RACHNA
Vidyaparibhaal

**Dr. O.P. Bhalla Central Library**
**Manav Rachna International Institute of Research and Studies**
*(A deemed-to-be-university under section 3 of UGC Act 1956)*

## SIMILARITY VERIFICATION CERTIFICATE

Title: DEVELOPMENT OF AN ADVANCED IMAGE CLASSIFICATION ALGORITHM WITH INTEGRATED CAPTIONING

Researcher: MONIKA CHAWLA

Registration Number: MRIIRSPHD23/01311

Supervisor: Dr. Rashmi Agrawal

Department: Computer Application

This is to certify that the above mentioned Short Synopsis/~~Thesis~~ was scanned for similarity index through iThenticate Software on 26/12/24 with submission ID 3618:74811098 and the similarity index was found to be 7 %. The complete report running into 20 Pages, 6324 Word count with File Size 295.0 KB/MB is being forwarded to the Research Supervisor/DRC/DRB for review.

Date: Dec 26, 2024

Rajesh Kumar
Dr. Rajesh Kumar
University Librarian
Dr. O.P. Bhalla Central Library
Manav Rachna International
Institute of Research & Studies
Faridabad

The complete report of the above Short Synopsis/Thesis has been reviewed and the similarity index is below the accepted norms as per the University Grants Commission (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations 2018. The Short Synopsis/Thesis may be considered for further processing

Research Scholar

Research Supervisor

# Monika Chawla

## DEVELOPMENT OF AN ADVANCED IMAGE CLASSIFICATION ALGORITHM WITH INTEGRATED CAPTIONING.docx

Manav Rachna International Institute of Research and Studies

### Document Details

Submission ID
trn:oid:::3618:74811000

Submission Date
Dec 26, 2024, 3:03 PM GMT+5:30

Download Date
Dec 26, 2024, 3:04 PM GMT+5:30

File Name
DEVELOPMENT OF AN ADVANCED IMAGE CLASSIFICATION ALGORITHM WITH INTEGRATED CAPT....docx

File Size
295.0 KB

20 Pages

6,324 Words

39,082 Characters

Excluding matches
less than 14 words

Dr. Rajesh Kumar
University Librarian
Dr. O.P. Bhalla Central Library
Manav Rachna International
Institute of Research & Studies
Faridabad

# Monika Chawla

## DEVELOPMENT OF AN ADVANCED IMAGE CLASSIFICATION ALGORITHM WITH INTEGRATED CAPTIONING.docx

Manav Rachna International Institute of Research and Studies

### Document Details

Submission ID
trn:oid:::3618:74811000

Submission Date
Dec 26, 2024, 3:03 PM GMT+5:30

Download Date
Dec 26, 2024, 3:04 PM GMT+5:30

File Name
DEVELOPMENT OF AN ADVANCED IMAGE CLASSIFICATION ALGORITHM WITH INTEGRATED CAPT....docx

File Size
295.0 KB

20 Pages

6,324 Words

39,082 Characters

# 7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

## Filtered from the Report

- Bibliography
- Small Matches (less than 10 words)

## Match Groups

**29** Not Cited or Quoted **7%**
Matches with neither in-text citation nor quotation marks

**0** Missing Quotations **0%**
Matches that are still very similar to source material

**0** Missing Citation **0%**
Matches that have quotation marks, but no in-text citation

**0** Cited and Quoted **0%**
Matches with in-text citation present, but no quotation marks

## Top Sources

6% Internet sources
5% Publications
0% Submitted works (Student Papers)

## Integrity Flags

**1 Integrity Flag for Review**

**Hidden Text**
42 suspect characters on 1 page
Text is altered to blend into the white background of the document.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

## Match Groups

- 29 Not Cited or Quoted 7%
  Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

## Top Sources

6% Internet sources
5% Publications
0% Submitted works (Student Papers)

Excluding matches less than 14 words

Dr. Rajesh Kumar
University Librarian
Dr. O.P. Bhalla Central Library
Manav Rachna International
Institute of Research & Studies
Faridabad

## Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

| | | |
|---|---|---|
| 1 | Internet | |
| doaj.org | | 1% |
| 2 | Publication | |
| Arsenii Senokosov, Alexandr Sedykh, Asel Sagingalieva, Basil Kyriacou, Alexey Me... | | 1% |
| 3 | Internet | |
| www.ncbi.nlm.nih.gov | | 1% |
| 4 | Internet | |
| arxiv.org | | 1% |
| 5 | Internet | |
| discovery.researcher.life | | 1% |
| 6 | Internet | |
| www.mdpi.com | | 0% |
| 7 | Internet | |
| bmcmedimaging.biomedcentral.com | | 0% |
| 8 | Internet | |
| web.archive.org | | 0% |
| 9 | Internet | |
| curin.chitkara.edu.in | | 0% |
| 10 | Internet | |
| cdn.techscience.cn | | 0% |

| | | |
|---|---|---|
| 11 | Internet | |
| mdpi-res.com | | 0% |
| 12 | Internet | |
| www.ijircst.org | | 0% |
| 13 | Publication | |
| Ajay Sharma, Pramod Kumar Mishra, "DRI-UNet: dense residual-inception UNet f... | | 0% |
| 14 | Publication | |
| Ru Wang, Yinju Nie, Jie Geng, "Multiscale Superpixel-Guided Weighted Graph Con... | | 0% |
| 15 | Publication | |
| Tiantao Xian, Zhixin Li, Tianyu Chen, Huifang Ma, "Dynamic Transformer for Imag... | | 0% |
| 16 | Internet | |
| eprints.leedsbeckett.ac.uk | | 0% |
| 17 | Internet | |
| link.springer.com | | 0% |
| 18 | Internet | |
| mis.itmuniversity.ac.in | | 0% |
| 19 | Internet | |
| shodhgangotri.inflibnet.ac.in | | 0% |
| 20 | Internet | |
| www.ijraset.com | | 0% |